

5_5_Magic_BLAST_contig_subsets

July 25, 2021

```
[1]: import numpy as np
import math
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
from matplotlib.ticker import MaxNLocator
from pandas.plotting import scatter_matrix
import pathlib
import warnings
warnings.filterwarnings('ignore')
```

```
[2]: from IPython.core.display import display, HTML
display(HTML("<style>.container { width:95% !important; }</style>"))
```

<IPython.core.display.HTML object>

```
[3]: PROJECT_CODE='PRJNA607174'
BASE_PATH = f'/mnt/1TB_0/Data/Assembly/{PROJECT_CODE}/'

dbname='nt'
kmer='k141'

f_contigs_file_tail=f'_{dbname}_magic_blast_asc_contigs.txt'
```

```
[4]: subsets=['vector','virus']
```

```
[5]: sra_list=['SRR11119759','SRR11119760','SRR11119761',\
'SRR11119762','SRR11119763','SRR11119764',\
'SRR11119765','SRR11119766','SRR11119767','SRR12053850']
```

```
[6]: def read_matched(asc_file, ignore_string=None):
    contigs=[]
    accessions=[]
    descriptions=[]
    cigars=[]
    with open(asc_file, 'r') as infile:
        data = infile.readlines()
        for i in data:
```

```

        output=i.split('\t')
        descr=output[2].split(' ',1)[1]
        if ignore_string is not None and ignore_string in descr:
            pass
        else:
            contigs.append(output[0])
            accessions.append(output[2].split(' ')[0])
            descriptions.append(descr)
            cigars.append(output[5])
    return contigs, accessions, descriptions, cigars

```

```

[7]: def process_file(asc_file, sra, ignore_string=None):
    contigs, accessions, descriptions, cigars=read_matched(asc_file,
    ↪ ignore_string)
    sra_list=[sra] * len(contigs)
    df = pd.DataFrame(list(zip(sra_list, contigs, accessions, descriptions,
    ↪ cigars)),
                        columns=['sra', 'contig', 'accession','description','cigar'])
    return df

```

```

[8]: def plot_df(df, dataset):
    df['sra'].value_counts().plot(kind='bar')
    plt.title(dataset + ' contig counts by SRA')
    plt.show()

```

```

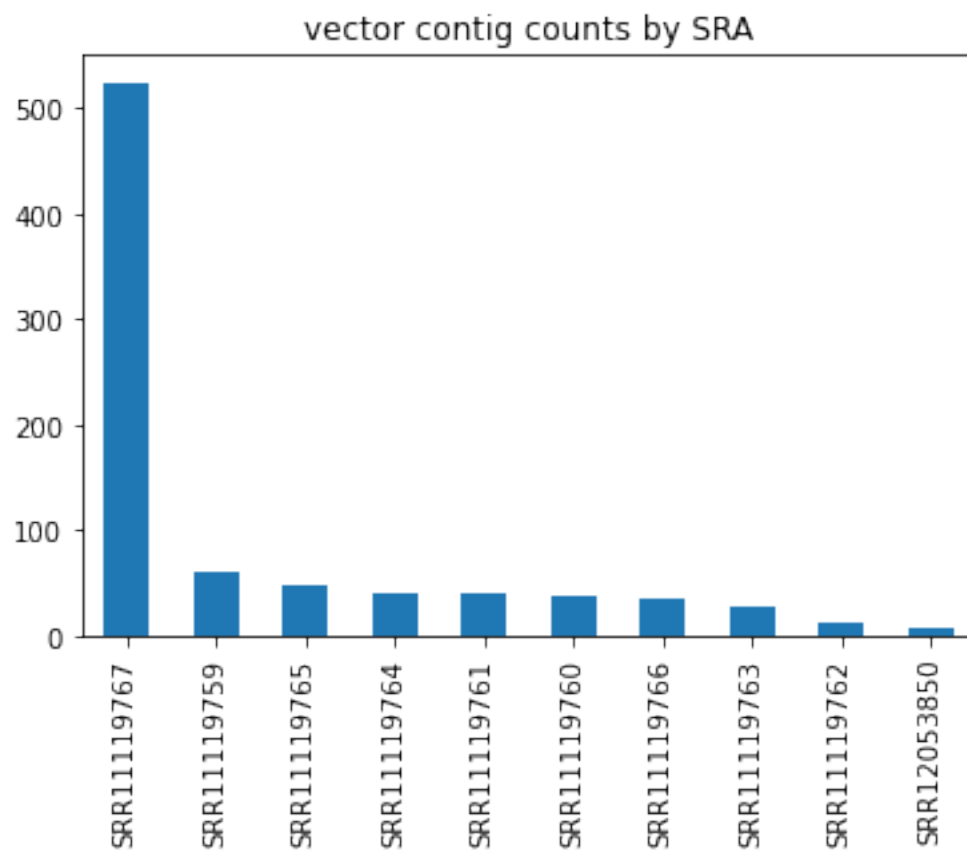
[9]: def workflow():
    for dataset in subsets:
        path = BASE_PATH+'./contig_subsets/'+dataset+'/'
        frames=[]
        for sra in sra_list:
            f=sra+'_'+dataset+'_subset'+ '_' +f_contigs_file_tail
            if dataset=='virus':
                df=process_file(path+f, sra, ignore_string='retrovirus')
            else:
                df=process_file(path+f, sra)
            if len(df)>0:
                frames.append(df)
        df_result = pd.concat(frames)
        df_result.to_csv(path+PROJECT_CODE+'_'+dataset+'_dataframe.csv')
        plot_df(df_result, dataset)

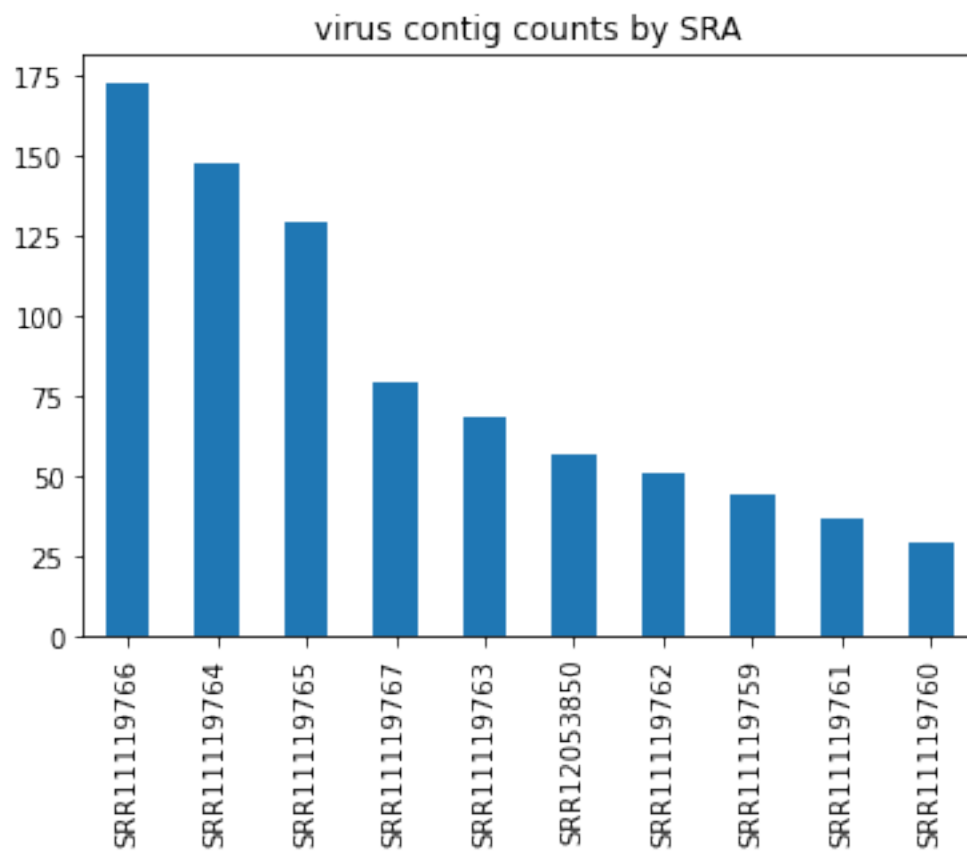
```

```

[10]: workflow()

```





[]:

[]: